International Journal on New Trends in Education and Their Implications
July 2023 Volume: 14, Issue: 1, ISSN 1309-6249

Investigation of Measurement Invariance of PISA Science Literacy Self-Efficacy Scale in Different Achievement Groups

Yusuf KASAP [1], Ömer Cem KARACAOĞLU[2]

Abstract

The aim of the study is to examine the measurement invariance of the science literacy self-efficacy items in the PISA 2015 student questionnaire in different science literacy achievement groups. In this context, the sample of the study consists of 5556 students from Korea representing the high achievement group, 6648 students from Spain representing the medium achievement group, and 5862 students from Turkey representing the low achievement group who responded to the Science self-efficacy questionnaire in PISA 2015. The self-efficacy model for science literacy, which was created by exploratory factor analysis in the first step with 18066 student data from three countries within the scope of the research, was confirmed with 8 items and a single factor. In the second step, it was examined whether the model provided measurement invariance in different success groups. In the process of examining the invariance, multi-group confirmatory factor analysis was performed, and the invariance of the model was examined gradually. The results of the research show that the self-efficacy model for science literacy provides shape invariance, metric invariance and scale invariance in different achievement groups. Since the model provides partial invariance between countries in different achievement groups, comparing the scores of individuals in the countries from the observed variables included in the model will give accurate results.

Keywords: PISA data, Science Literacy, Self-efficacy Scale, Measurement İnvariance

Recommended Citation: Kasap, Y. & Karacaoglu, O. C. (2023). Investigation of measurement invariance of PISA science literacy self-efficacy scale in different achievement groups, *International Journal on New Trends in Education and Their Implications (IJONTE),* 14(1), p. 88-98.

Introduction

In the field of education, measurement and evaluation processes have an important place in order to determine the learning levels of individuals. There are many types of exams and assessment types in the assessment and evaluation. These exams can be done both nationally and internationally. One of the exams held in this context is the International Student Assessment Program (PISA- Programme) conducted by the Organization for Economic Cooperation and Development (OECD). for International Student assistance. The purpose of PISA, which Turkey has been participating in since 2003 and which is carried out every three years by focusing on specific subject areas, is to determine the achievement levels of 15-year-old individuals in three areas: mathematical literacy, science literacy and reading skills, and to determine whether they can transfer the skills they have acquired in these areas to daily life. In addition, the results obtained from PISA allow comparisons between participating countries and different groups.

PISA science literacy self-efficacy scale is a scale used to measure students' self-confidence in science subjects (Latifah et al., 2019). Examining the measurement invariance of this scale in different

[1] Dr., Ankara, Turkey, yusufkasap5806@hotmail.com, ORCID: https://orcid.org/) 0000-0002-5114-1175

[2] Assoc. Prof. Dr., Izmir Turkey, cemkaracaoglu@gmail.com, ORCID: https://orcid.org/ 0000-0003-2474-5106

achievement groups can provide important information about the validity and reliability of the scale. Measurement invariance means that the property that a scale measures is measured in the same way among different groups to which the scale is applied (Emerson et al., 2017). Ensuring measurement invariance is an indication that the scale is valid and reliable (Ireland et al., 2019). In order to examine the measurement invariance between different achievement groups, first of all, students should be divided into different achievement groups. This distinction is usually made based on students' scores on the PISA exam. Then, a sample is selected from each achievement group and the scale is applied. In order to examine the measurement invariance of the scale between different achievement groups, methods such as internal consistency coefficients of the scale, factor analysis, and item analysis of the scale can be used (Jovanović et al., 2020). These methods show that the scale measures the same across different groups. Examining the measurement invariance of the PISA science literacy self-efficacy scale between different achievement groups can provide important information about the validity and reliability of the scale. This information can be taken into account in the process of developing and improving the scale.

Comparisons of the psychological structures discussed in the studies are also made between different groups. Construct validity is related to the extent to which the measurement tool can reveal a psychological structure or feature (Anastasi, 1988). Due to the differentiation of individuals and therefore groups from each other, the results obtained from the measurements may differ. In addition to the differences between the individual and the group, there are also different factors affecting the measurement results. Therefore, since the item and test statistics calculated in the context of validity and reliability are calculated depending on the group, they may show different results depending on the characteristics of each group (Crocker & Algina, 1986). In addition to these factors, the instrument measuring a defined psychological structure should be made meaningful for all groups while making intergroup comparisons such as gender differences and regional differences in measurement processes between groups. Before intergroup comparisons, measurement invariance should be considered so that it can be determined that the same construct or constructs are being measured in each of the groups in the comparison. Measurement invariance is that the relationship between latent variables and observed variables is the same between groups (Widaman & Rice, 1997). On the other hand, measurement invariance is a prerequisite for making comparisons between different groups (Horn & McArdle, 1992). For example, while investigating the job satisfaction level of groups of American and Chinese workers, the data to be obtained from the scale used will vary according to the way the groups interpret the items (Cheung & Rensvold; 1999). When making comparisons between groups in studies in the field of social sciences, the assumption that there is no difference due to the measurement tool may weaken the study. For this reason, the invariance of the measurement tool should be revealed by the analyzes made. Measurement invariance emerges to make comparisons between different groups (Cheung & Rensvold; 2000) and its basis lies in ensuring the validity of the measurements made for the comparison groups (Tyson, 2004). Ensuring the cross-cultural validity of the scales used ensures the generalizability of the scales and the reliability of the results obtained (Marsh et al., 2006). When the relevant literature is examined, it is seen that measurement invariance studies are carried out especially to determine the cross-cultural generalizability of psychological constructs and the comparability of these constructs. Studies on the generalizability and comparability of the mentioned structures are more about the measurement tool than the individuals in the sample, and it aims to determine whether the results obtained from the measurement tool reveal an equal structure.

The contributions of studies on measurement invariance (Ayvalı, 2016; Akyıldız, 2009; Bahadır, 2012; Başusta & Gelbal 2015; Kıbrıslıoğlu, 2015; Uyar & Doğan, 2014) to the literature are very valuable. In addition to the studies on PISA, it has been reported that measurement invariance studies conducted on large-scale exams such as TIMSS and with individual measurement tools do not provide measurement invariance between different language, culture or gender groups (Ercikan & Koh, 2005; Önen, 2009; Steinmetz et al., 2009; Wu et al., 2007). However, some studies with different samples have revealed that the PISA science literacy self-efficacy scale provides measurement invariance between different achievement groups (Odell et al., 2021; Güngör & Kabasakal, 2020; Uysal & Arıkan, 2018).

In order to examine measurement invariance, it requires four steps and hypothesis testing. These are;

Shape invariance: It is the initial level of measurement invariance. At this stage, hypothesis testing is performed, indicating that the factor structure of the measurement tool is invariant between groups. If there is evidence for stylistic invariance, it means that the items of the measurement tool measure similar psychological structure between groups.

Metric invariance: If the responses of separate groups to the items are similar, it is possible with metric invariance that the scores obtained from separate groups are comparable (Steenkamp & Baumgartner, 1998). At this stage, the hypothesis that the regression coefficients for the items that make up the measurement tool, also known as the factor loads, is invariant between the groups is tested.

Scalar invariance: At this stage, the hypothesis that the constant coefficient in the regression equation created with the items constituting the measurement tool is invariant between the groups is tested. For scale invariance to exist, both metric invariance and regression require the same origins in the equation. This stage of measurement invariance; means that the value on the implicit structure is equal to the value on the observed value.

Strict invariance: At this stage, the hypothesis that the errors related to the items that make up the measurement tool is invariant among the groups is tested.

Each of these levels; It is created based on the model in the previous stage. Thus, measurement invariance at any level is based on examining the level of fit of the model at that level and the model at the previous level to the data.

Generally, it is not possible to provide all of the invariance stages. In such cases, according to the purpose of the scale used, partial invariance is also considered as a degree of measurement invariance. In order to be able to compare the means of the scale in the groups, it is necessary to ensure invariance up to scalar (full score equivalence) invariance (Dimitrov, 2010).

Structural equation model studies are more focused on developing theory. While developing the theory, studies are carried out to determine the existing relationships between the variables studied. Structural equation model applications to determine measurement invariance give generalizable information about the theory that is tried to be put forward. In this way, it can be said that measurement invariance studies have a very important role in determining the structures that are tried to be defined. Constancy is directly related to the construct validity of measurement tools. Invariance studies, especially applied for large-scale exams, give valid results for comparing groups. As a result, it is thought that the intensification of invariance studies by considering various variables for various measurement tools will contribute to the literature.

The self-efficacy scale is a psychological tool used to measure the level of self-confidence of individuals. This scale can give healthier results by examining factors such as measurement invariance and equality of self-efficacy items in different achievement groups. The purpose of examining measurement invariance in different achievement groups is to verify that the self-efficacy scale is valid for all groups (Özgen & Bindak, 2008). The purpose of examining the equality of self-efficacy items is to verify that each item of the scale is equally meaningful for different groups. Therefore, each item of the PISA 2015 science self-efficacy scale should be examined to ensure that it has the same meaning for different groups. In order for the science self-efficacy scale to be used as a valid and reliable tool, examining the measurement invariance and equality of self-efficacy items in different achievement groups will shed light on the comparisons and researches to be made. Structural equation modeling can be used to examine the invariance of the science self-efficacy scale and the equality of the items and to generalize the results (Kaynak, 2012; Şimşek, 2020).

The aim of the study is to examine the measurement invariance of the PISA 2015 science literacy self-efficacy scale in the achievement groups of Korea, Spain and Turkey, and the equality of the self-efficacy items. For this purpose, For this purpose, it will be ensured to test the measurement invariance of self-efficacy for science literacy of students participating in the PISA 2105 exam for countries with different achievement levels. In this context, it is thought that the study will provide more valid results on the significance of the comparisons of self-efficacy items for science literacy in

PISA 2015 according to countries with different achievement levels. Testing the measurement invariance in the study; It is thought that any comparison made on the basis of countries with different success levels will be a reference to other studies that will make it meaningful. In addition, since studies on measurement invariance in the literature mostly focus on different language, different culture and gender groups, it is thought that the measurement invariance study for different success groups in this research will contribute to the literature. The research problem is whether there is evidence of item measurement invariance for countries with different achievement levels by using the scale data measuring self-efficacy for science literacy in PISA 2015.

Method

Research Design

In this study, the equality of the items measuring self-efficacy for science literacy in the PISA 2015 application was examined in terms of Korea, Spain and Turkey. In this direction, this study is a study in the relational survey model.

Universe and Sample

For the purpose of the study, countries; purposive sampling method was used because it was determined according to low, medium and high achievement levels, taking into account the OECD PISA science literacy average. Because purposeful sampling is the selection of a sample rich in information in line with the purpose of the study in order to conduct in-depth research (Büyüköztürk et al., 2018).

For the purpose of this study, one country each for the lower, middle and upper groups was selected from the 72 countries participating in the PISA 2015 exam. For the selection of the countries, the average of each group was determined and the country closest to this average in each group was selected. These three selected countries; Korea, which has the same average with the average of the countries with achievement above the OECD science average in the upper group, Spain, which has the same average with the average of the countries with achievement not significantly different from the OECD science average in the middle group, and Turkey, which has the same average with the average of the countries with achievement below the OECD science average in the lower group, constitute the study sample. The study sample consists of students participating in the PISA 2015 application from Korea, Spain and Turkey. In the research, student data from Korea (5556), Spain (6648) and Turkey (5862) were used.

Data Collecting

Research data is taken from http://www.oecd.org/pisa/data/2015database/. These data were shared in 2016. The data in the study were obtained from the self-efficacy items for science literacy in the ST129 coded student questionnaire of the PISA 2015 application. Ethical principles were followed in obtaining the data.

Data Collection Tools

In this study, data belonging to the 8-item science literacy self-efficacy subscale of the student questionnaire coded ST129 and applied in PISA 2015 were used. This scale is a 4-point Likert- type structure that scales between strongly disagree (1) strongly agree (4). The code and item form of the items in the ST129 coded student questionnaire are given in Table 1.

Table 1

Items of the ST 129 Coded Student Questionnaire

Item Code	Materials
ST129Q01TA	Recognise the science question that underlies a newspaper report on a health issue.
ST129Q02TA	Explain why earthquakes occur more frequently in some regions than in others.
ST129Q03TA	Describe the role of antibiotics in the treatment of disease.
ST129Q04TA	Identify the science question associated with the disposal of garbage.
ST129Q05TA	Predict how changes to an environment will affect the survival of certain species.
ST129Q06TA	Interpret the scientific information provided on the labeling of food items.
ST129Q07TA	Discuss how new evidence can lead you to change your understanding about the possibility of life on Mars.
ST129Q08TA	Identify the better of two explanations for the formation of acid showers.

Data Analysis

SPSS23 and Mplus7 for Windows software were used to analyze the data. First of all, the missing data were examined and it was determined that the rate of missing data was below 1% and was randomly distributed. Since the item scores were at the level of the ranking scale, the median value was assigned instead of the missing data. In addition, for the exploratory factor analysis, it was examined whether the data showed a normal distribution and whether there was a multicollinearity between the scale items. According to the results obtained, it was determined that the data were normally distributed and there was no multicollinearity problem between the scale items. Then, after KMO and Barlett tests were performed to determine whether the data before the analysis were suitable for exploratory factor analysis, it was determined that the data set was suitable for exploratory factor analysis. Finally, exploratory factor analysis was performed and a measurement model was established according to the results of this analysis.

Multi-group confirmatory factor analysis (CGFA) was applied to test the model-data fit. In testing the model data fit, CFI, χ^2/df χ^2/df ratio and mean square error of estimation (RMSEA) indices were examined. In structural equation modeling analyzes, it was observed that chi-square values were often significant due to large samples (Brown, 2006). For a good model-data fit, it has been deemed appropriate to use RMSEA values and CFI values, which are stronger against statistical inadequacies caused by sample size and are more appropriate in large samples than other fit statistics (Cheung & Rensvold, 2002). It is accepted that a CFI value higher than 0.90 and a RMSEA value lower than 0.08 indicate a good model-data fit.

With multiple group applications, it is tried to determine whether the latent traits to be measured are different between the groups. In these applications, differences between confirmatory coefficients of fit (CFI) are used. There are studies suggesting that the CFI value is more appropriate as an alternative to chi-square when there are statistical inadequacies caused by the large sample size in determining the fit. Wu et al., (2007) explained that the difference between the CFI fit coefficients should be used instead of the chi-square difference in their study with TIMSS data as follows. The chi-square difference is a function of the sample size, and it is not correct to use it alone in the decision for measurement invariance.

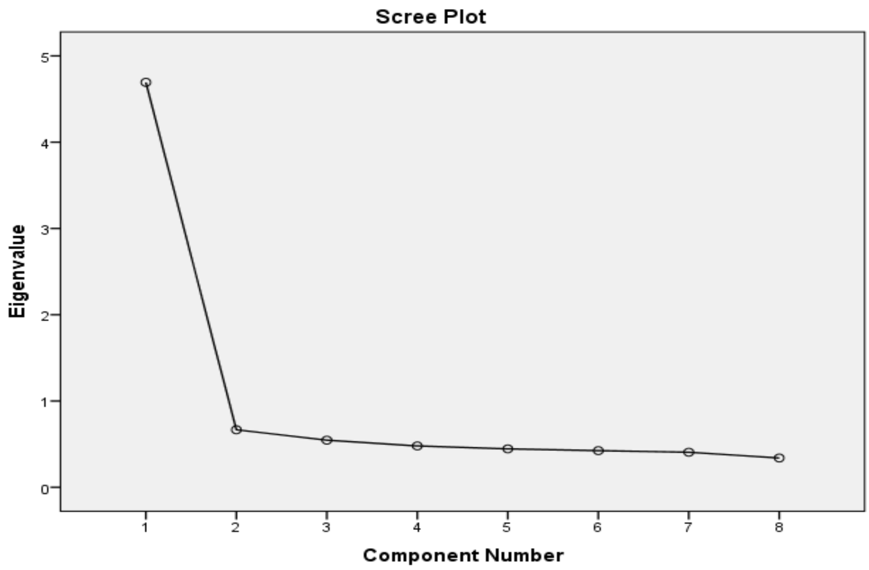
As an invariance test in the model used in this study, CFI differences (ΔCFI) obtained from the application of shape invariance, metric invariance, scale invariance and strict invariance were compared. The main reason for examining ΔCFI values is that the fit coefficients provide information about the relationship between latent scores and observed scores. In order to fulfill the condition of invariance, the invariance was examined by taking into account that ΔCFI values were in the range of $0.01 \geq \Delta CFI \geq -0.01$. Furthermore, comparisons were made based on ΔCFI values as well as SRMR difference values (ΔSRMR) for metric and strict invariance. Acceptance levels for these comparisons are $\Delta CFI \geq -.010$ and $\Delta SRMR \geq .015$ for metric invariance when $n>300$, and $\Delta CFI \geq -.010$ and $\Delta SRMR \geq 0.010$ for strict invariance. On the other hand, when $n<300$, $\Delta CFI \leq -.005$ and $\Delta SRMR \geq .025$ conditions must be met for metric invariance, while $\Delta CFI \geq -.005$ and $\Delta SRMR \geq 0.005$ conditions must be met for strict invariance.

Findings

Exploratory factor analysis was performed on 8 items in the PISA self-efficacy questionnaire, and it was checked whether the item clusters were suitable for the theory and model. The eigenvalues of the factors scree plot chart is given below.

Figure 1

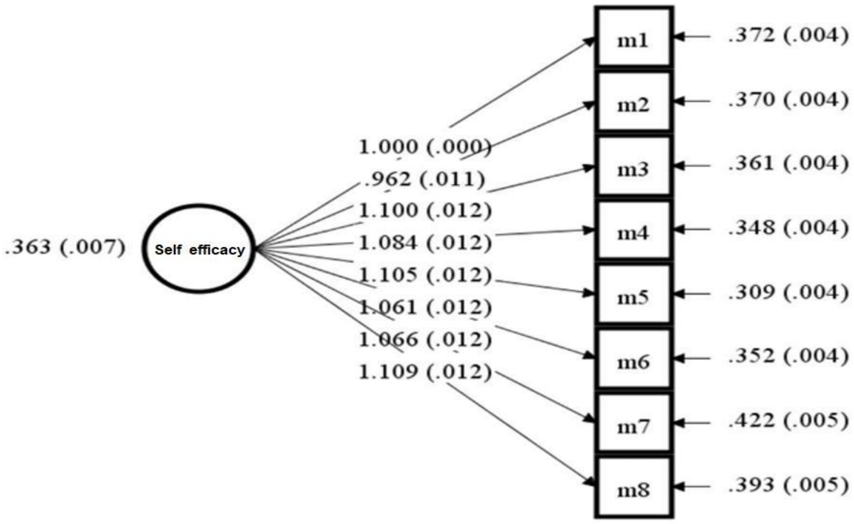
Figure 1: Scree plot drawn depending on the eigenvalues of the factors



In Figure 1, it is seen that 1 factor is dominant in the scree plot drawn regarding the eigenvalues of the factors. Since 8 items to which exploratory factor analysis was applied measure self-efficacy towards science, all of the items were gathered under 1 factor. The total variance explained by a factor is 58.662%. The measurement model created with the path diagram obtained by exploratory factor analysis is as in Figure 2.

Figure 2

Path diagram for the shape invariance obtained from all data



The fit statistics for the model in Figure 2 were found to be at an acceptable level of fit. Since the sample size is very large, CFI, RMSEA and SRMR values were taken into account. Considering the data's fit criteria for the model (CFI>0.90, RMSEA<0.08 and SRMR<0.08), the data used was found to be compatible with the single-factor model since CFI=0.967, RMSEA= 0.080 and SRMR=0.026.

Information on the tested invariance levels is in Table 2. Shape invariance model in the table; factor loadings, factor correlations and error variances free model, metric invariance model; constant factor loadings, factor correlations and error variances free model, scale invariance model; factor loadings and factor correlations constant, error variance free model and strict invariance model; factor loads, factor correlations and error variances represent the fixed model.

The model in Figure 2 shows the model with free factor loadings, factor correlations and error variances. The fit statistics and the difference values of the fit statistics calculated when restrictions are imposed on the model without restrictions are as shown in Table 2.

Table 2

Fit statistics for measurement invariance levels

Stages	χ2	df	RMSEA	SRMR	ΔSRMR	CFI	ΔCFI
Shape invariance	5604,563	75	0.111	0.055	-	0.925	-
Metric invariance	5692,472	82	0.107	0.058	0.003	0.924	-0.001
Scale invariance	6236.712	88	0.108	0.061	0.006	0.917	-0.008
Strict invariance	11486.807	106	0.134	0.136	0.081	0.846	-0.079

If comments are made for each level in terms of invariance according to the indexes given in the table; When the fit statistics for shape invariance are examined (CFI>0.90, and SRMR<0.08), it is seen that the model adapts to the data in the shape invariance stage. When the ΔCFI (CFI difference value) fit index was examined to decide whether the shape invariance was achieved or not, it was decided that this stage was approved. Confirmation of figural invariance indicates that the measured structure is similar from group to group, that is, students from Korea, Spain and Turkey have similar conceptual perspectives in responding to scale items. When the fit statistics for metric invariance are examined (CFI>0.90, and SRMR<0.08), it is seen that the model adapts to the data in the metric invariance stage. Obtained from CFI and SRMR difference test to decide metric invariance (ΔCFI and ΔSRMR) values were analyzed and interpreted. The findings show that the relationship between the property measured by the items of the self-efficacy scale and the self-efficacy dimension is similar for students from Korea, Spain, and Turkey. When the fit statistics for scale invariance are examined (CFI>0.90, and SRMR<0.08), it is seen that the model adapts to the data in the scale invariance stage. When the ΔCFI value calculated with the CFI fit indices obtained from the multi-group DFA application was examined to determine the scale invariance, it was determined that there was no bias on the basis of the items for students from Korea, Spain and Turkey, since the obtained value met the condition of invariance. As a result, the hypothesis that the constant coefficients in the regression equations established for the items are invariant for students from Korea, Spain and Turkey has been confirmed. At the last level, when the fit statistics for strict invariance were examined, it was determined that the model did not fit the data in the strict invariance stage, since it did not meet the condition (CFI<0.90 and SRMR>0.08). In addition, the hypothesis that the errors for the items in the measurement tool are invariant between the groups compared, was calculated using the fit indices (ΔCFI and ΔSRMR) obtained from the multi-group DFA. not validated considering its value. Thus, it was concluded that the errors regarding the items in the measurement tool differed between the groups compared. As a result, all of the other types of invariance are satisfied except strict invariance. Thus, partial invariance of the self-efficacy scale was provided for Korea, Spain and Turkey. As a result of the findings obtained from all analyses, it will be meaningful to compare the averages of self-efficacy for the countries with different success levels regarding the model created by considering the EFA results.

Discussion, Conclusion, and Recommendations

In this research to determine the measurement invariance for countries; Within the scope of the measurement model created for science literacy self-efficacy; Shape invariance, metric invariance, and scale invariance are confirmed. However, strict invariance has not been confirmed. According to this result, only partial invariance of the scale was achieved. The results of the measurement invariance applications made within the framework of the research show that the psychometric structure obtained from the measurement model, which consists of the items in the PISA student questionnaire and prepared to reveal the students' self-efficacy towards science, can be generalized for three countries, namely Korea, Spain and Turkey. This result is similar to other studies revealing that the PISA science literacy self-efficacy scale provides measurement invariance between different regional and gender groups (Ding et al., 2023; Güngör & Kabasakal, 2020; Uysal & Arıkan, 2018).

The findings of the study provide evidence that the PISA self-efficacy questionnaire is unbiased with respect to the countries studied and provides valid and reliable results in revealing the characteristics of Korean, Spanish and Turkish students' self-efficacy constructs. This finding is supported by the finding that the PISA science literacy self-efficacy scale is valid and statistically reliable (Mohd Dzin & Lay, 2021). Ding et al. (2023) analyzed the data of over 600,000 students from 80 countries and two different PISA assessments, and found that there was a large degree of invariance in self-efficacy factors and that factor tools could not be compared across all participating countries. In the study, it was stated that as the number of groups and sample size increased, the possible violation of invariance increased. Similarly, Rutkowski & Svetina (2014) stated that a large group is not suitable in the context of various sample sizes or should be adjusted especially when the number of groups is large. These findings show that the selection of Korea, Spain and Turkey, which are in different rankings instead of more countries in the research, is correct. Similarly, Uysal and Arıkan (2018) argued that the science self-efficacy scale met all stages of gender invariance in both 2006 and 2015, and provided measurement invariance between gender and country groups.

It can be said that the measurement model, in which measurement invariance is investigated, measures similarly for Korean, Spanish and Turkish students, and thus, based on the results obtained from the measurement tool, it can be said that it would be appropriate to compare Korean, Spanish and Turkish students within the scope of the measurement model studied.

References

Akyıldız, M. (2009). Cross-country comparison of construct validity of the PIRLS 2001 test. *The Journal of Yüzüncü Yıl University, Faculty of Education.* one, 18-47. https://dergipark.org.tr /en/pub/yyuefd/issue/13711/165993

Ayvali, M. (2016) *Investigation of measurement invariance of PISA 2012 Mathematical Literacy Test* [Master Thesis]. Mediterranean University.

Anastasi, A. (1988). *Psychological testing.* Macmillan Publishing Company.

Bahadir, E. (2012). *Analysis of the variables affecting the reading skills of students in Turkey according to the Program for International Student Assessment (PISA 2009), according to the regions* [Master Thesis]. Hacettepe University.

Başusta, N.B. & Gelbal, S. (2015). Testing measurement invariance in comparisons between groups: PISA student questionnaire example. *Hacettepe University Faculty of Education Journal,* 30(4), 80-90. http://efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/1773-published.pdf

Büyüköztürk, Ş., Kılıç-Çakmak, E., Akgün, Ö.E., Karadeniz, Ş., & Demirel, F. (2018). *Scientific Research Methods.* (25th Edition). Pegem Publishing.

Brown, T.A. (2006). *Confirmatory factor analysis for applied research.* Guilford press.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory. Toronto:* Holt, RineHart, and Winston, Inc.

Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural equation modeling: a multidisciplinary journal,* 14(3), 464-504. https://doi.org/10.1080/10705510701301834

Cheung, G.W. & Rensvold, R.B. (1999). Testing factorial invariance across groups: a reconceptualization and proposed new method. *Journal of Management,* 25(1), 1–27. https://doi.org/10.1177/014920639902500101

Cheung, G.W. & Rensvold, R.B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of cross-cultural,* 31(2), 187-212. https://doi.org/10.1177/00220 *psychology22100031002003*

Cheung, G.W. & Rensvold, R.B. (2002). Evaluating goodness -of-fit indexes for testing measurement invariance. *Structural equation modeling,* 9(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Dimitrov, D.M. (2010). testing for factorial invariance in the context of construct validation. *Measurements and Evaluation in Counseling and Development,* 43(2), 121-149. https://doi.org/10.1177/0748175610373459

Ding, Y., Yang Hansen, K., & Klapp, A. (2022). Testing measurement invariance of mathematics self-concept and self-efficacy in PISA using MGCFA and the alignment method. *European Journal of Psychology of Education*, 1-24. https://doi.org/10.1007/s10212-022-00623-y

Emerson, S.D., Guhn, M., & Gadermann, A.M. (2017). Measurements invariance of the Satisfaction with Life Scale: Reviewing three decades of research. *Quality of Life Research,* 26, 2251-2264. https://doi.org/10.1007/s11136-017-1552-2

Ercikan, K. & Koh, K. (2005). Examining the construct comparability of the English and French versionf of TIMSS. *International Journal of Testing, 5,* 23-35. https://doi.org/10.1207/s15327574ijt0501_3

Güngör, M. & Kabasakal, K.A. (2020). Investigation of measurement invariance of science motivation and self- efficacy model: PISA 2015 Turkey sample. *International Journal of Assessment Tools in Education, 7* (2), 207-222. https://doi.org/10.21449/ijate.730481

Horn, J.L. & McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. *experimental aging Research,* 18(3), 117-144. https://doi.org/10.1080/03610739208253916

Ireland, M.J., Day, J.J., & Clough, B.A. (2019). Exploring scale validity and measurement invariance of the Toronto Mindfulness scale across levels of meditation experience and proficiency. *Journal of clinical psychology, 75* (3), 445-461. https://doi.org/10.1002/jclp.22709

Jovanovic, V., Lazić, M., & Gavrilov-Jerković, V. (2020). Measuring life satisfaction among psychiatric patients: Measurement invariance and validity of the Satisfaction with Life Scale. *Clinical Psychology & Psychotherapy, 27* (3), 378-383. https://doi.org/10.1002/cpp.2434

Source, Z.N. (2012). *Structural equation models* [Master's thesis]. Istanbul Commerce University.

Kıbrıslıoğlu, N. (2015). *Investigation of measurement invariance of PISA 2012 mathematics learning model according to cultures and gender* [Master's Thesis]. Hacettepe University.

Latifah, S., Susilowati, N. E., Khoiriyah, K., & Rahayu, R. (2019, February). Self-efficacy: Its correlation to the scientific-literacy of prospective physics teacher. In *Journal of Physics: Conference Series* (Vol. 1155, No. 1, p. 012015). IOP Publishing. https://10.1088/1742-6596/1155/1/012015

Marsh, H.W., Hau, K.T., Artelt, C., Baumert, J., & Peschar, J.L. (2006). OECD's brief self- report measure of education psychology's most useful affective constructs: Cross- cultural, psychometric comparisons across 25 countries. *International Journal of Testing,* 6(4), 311-360. https://doi.org/10.1207/s15327574ijt0604_1

Mohd Dzin, N.H., & Lay, Y.F. (2021). Validity and reliability of adapted self- efficacy scales in Malaysian context using PLS-SEM approach. *Education Sciences,* 11(11), 676. https://doi.org/10.3390/educsci11110676

Odell, B., Gierl, M., & Cutumisu, M. (2021). testing measurement invariance of PISA 2015 mathematics, science, and ICT scales using the alignment method. *Studies in Educational Evaluation, 68,* 100965. https://doi.org/10.1016/j.stueduc.2020.100965

Onen, E. (2009). *Investigation of measurement invariance with structural equation modeling techniques* [PhD Thesis]. Ankara university.

Ozgen, K. & Bindak, R. (2008). Developing a mathematical literacy self-efficacy scale. *Kastamonu Journal of Education, 16* (2), 517-528. https://dergipark.org.tr /en/pub/kefdergi/issue/49100/626538

Rutkowski, L. & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large scale international surveys. *educational and psychological measurement, 74* (1), 31-57. https://doi.org/10.1177/0013164413498

Steenkamp E.M. & Baumgartner H. (1998). Assesing measurement invariance in cross-national consumer research. *Journal of Consumer Research,* Volume 25, Issue 1, June 1998, Pages 78–90, https://doi.org/10.1086/209528

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S.H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups of people values measurement. *Qual Quant, 43,* 599-616. https://doi.org/10.1007/s11135-007-9143-x

Şimşek, O.F. (2020). *Introduction to structural equation modeling: Fundamentals and applications of LISREL.* Equinox

Tyson, H.E. (2004). Ethnic differences using behavior rating scales to assess the mental health of children: a conceptual and psychometric critique. *Child Psychiatry and Human Development,* 34(3), 167-201. 11th. https://doi.org/10.1023/B:CHUD.00000014996.26276.a5

Uyar, S. & Doğan, N. (2014). Investigation of measurement invariance of learning strategies model in different groups in PISA 2009 Türkiye sample. *International Journal of Turkish Educational Sciences, 2,* 30-43. https://dergipark.org.tr /en/pub/goputeb/issue/33496/380021

Uysal, N.K. & Arıkan, C.A. (2018). Measurements invariance of science self- efficacy scale in PISA. *International Journal of Assessment Tools in Education,* 5(2), 325-338. https://doi.org/10.21449/ijate.379508

Wang, Y.L., Liang, J.C., & Tsai, C.C. (2018). cross cultural comparisons of university students ' science learning self- efficacy: structural relationships among factors within science learning self- efficacy. *International Journal of Science Education,* 40(6), 579-594. https://doi.org/10.1080/09500693.2017.1315780

Wu, D.A., Li, Z., & Zumbo, B.D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: a demonstration with TIMSS data. *Practical Assesment, Research & Evaluation,* 12(3), 1-26. https://doi.org/10.7275/mhqa-cd89

Widaman, K.F. & Reise, S.P. (1997). Exploring the measurement invariance of psychological instruments: applications in the substance use domain. *The science of prevention: Methodological advances from alcohol and substance abuse research,* 281-324. https://doi.org/10.1037/10222-009